Predicting The Road Traffic Density Based on Twitter Using The TR-P Method

Arief Wibowo[1,2], Edi Winarko[1], Azhari[1]
[1)]Department of Computer Science and Electronics, Universitas Gadjah Mada, Indonesia
[2)]Department of Information System, Faculty of Information Technology, Universitas Budi Luhur, Indonesia

Summary
This study contains work activities in order to build models of prediction system of traffic density based on twitter text data. The Traffic Road-Prediction (TR-P) method we have proposed consists of extracting and mining text data on road traffic. Based on the patterns that have been found from the learning process, we build a prediction model of road traffic density using text mining approach. The result of learning model validation using k-fold method has an accuracy of 96.31%, while the highest model testing accuracy reaches 91.34%. The road traffic density conditions of the proposed system have been visualized into several color classes of road traffic density, such as red, orange, yellow and green. This visualization makes it easier for users to understand the level of road traffic density resulting from predicted systems that have been built.

Key words:
Text Classification, Data Mining, Road Traffic Density Prediction, TR-P Method.

1. Introduction

Twitter is a very popular social media. Various information is submitted via Twitter by users, individually or by the organization for their respective purposes. Twitter messages (Tweets) can be easily read, even accessible in relatively large numbers over a period. The ease and practicality of collecting tweet archives can be used for research, including the study of the road traffic prediction.

Various studies have used twitter as a research resource in the field of traffic, such as [1]–[4]. Previous studies also used Twitter to predict the current status of traffic flow [5]–[8]. Nevertheless, there has not been made a study yet to predict future traffic density using Twitter.

This study used twitter data as a source data to build models capable of predicting future road traffic density. The difference in our study is the discovery of a new method called *"TR-P"* used to predict road traffic density conditions that will occur in the next few minutes, instead of the current time. An overview of predictive model development using TR-P method is shown in Fig. 1.

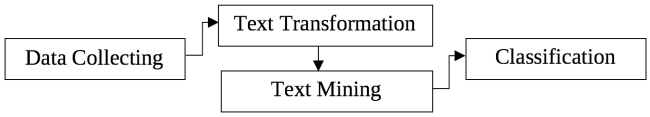

Fig. 1: The TR-P method frameworks

Fig. 1 shows that the TR-P method consists a series of text processing tasks starting from the collection of Twitter data, followed by processing activities in the form of transformation and text mining to obtain rules or patterns about the data. The rules are used to classify text data as a prediction model of road traffic density.

2. Text Processing

This research data source uses a collection of archive tweets about road traffic in Jakarta. Tweets are collected from the community of road users such as drivers or pedestrians who have sent messages about road traffic congestion conditions through police twitter accounts that deal with traffic management (Traffic Management Center of the Polda Metro Jaya as known as TMCPoldaMetro). Tweets of research data sources collected over a period of one year (2015-2016) with examples that can be seen in Table 1.

Table 1: The TMCPoldaMetro's Tweets

Date/Time	Twitter Messages
Wed Jan 14 03:43:43 2015	*10.37 Imbas Truk mengalami gangguan di flyover Tomang arah Kbn Jeruk lalu lintas padat di sekitar lokasi. @adriationo*
Thu Jan 15 03:25:37 2015	*10:33 #Lalin di sekitar depan UTC Koja terpantau padat merayap (Photo:@hkerbala)*
Fri Jan 16 11:22:43 2015	*18:19 Tol Dalam Kota Tomang arah Semanggi maupun arah sebaliknya padat RT @adit_aking*
Wed Jan 28 13:13:01 2015	*20:10 Lalu lintas dari arah Tomang menuju Harmoni terpantau padat begitu juga sebaliknya*
Wed Jan 28 22:16:25 2015	*05:15 Kecelakaan truk tabrak trotoar di dekat TL Halim Jaktim dan sudah dalam penanganan petugas #Polri*

In Table 1, it shows that the Tweets still requires further processing, such as word separation, text extraction, and analysis to form rules of attribute values that can be further processed.

Manuscript received August 5, 2017
Manuscript revised August 20, 2017

The pre-processing text is the task of preparing the Tweets for the next task as is the work in previous studies. The steps are data cleansing, tokenization, and word placement into each prepared attribute column [7]–[11]. The pre-processing task phases in this study includes the data cleaning and transformation phase of text into better data forms before the extraction process, include:

a. URL Removal
b. Illegal characters removal
c. Text standardization
d. Abbreviations transformation
e. Date and time extraction
f. Time zone adjustment

The pre-processing of text activities in this study is the task of data extraction tweet using the *"xTRoad"* method. The *"xTRoad"* is a text extraction method that is able to detect tweets about road traffic conditions in a particular or specific region accurately. Often on traffic tweet data, similar road names have been found in several different cities. In that case, the *"xTRoad"* method has been able to recognize roads that are specifically located in a town. The work scheme of the *"xTRoad"* extraction method can be seen in Fig. 2.

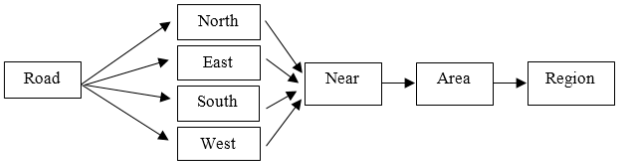

Fig. 2: The *xTRoad* extraction method

In the Fig. 2, it shows a road attribute (road name) is related to another path or area on the surrounding side. The road or related area may be on the north, east, south or west side. By connecting the road data or the area around it, it can be seen that the tweet contains road data in a particular city. The *"xTRoad"* method we proposed in this extraction process has successfully separated 2993 rows of data for the learning process in the next stage. The extracted tweets are contained the national road information in the city of Jakarta, which is the object boundary of this study.

3. Traffic Density Prediction Modeling

The main stage in this research is the model development for road traffic density prediction. We have chosen a data mining approach for text processing using classification techniques. Classification is a predictive data mining task aims at predicting (forecasting) a future value or unknown value of a dependent variable [12]. The reason we have chosen the classification rule is therefore very suitable for use in natural language based text processing to obtain the patterns, as previous researches [13]–[15]. The text classification activity scheme that becomes an element in the traffic density prediction system in this study is shown in Fig. 3.

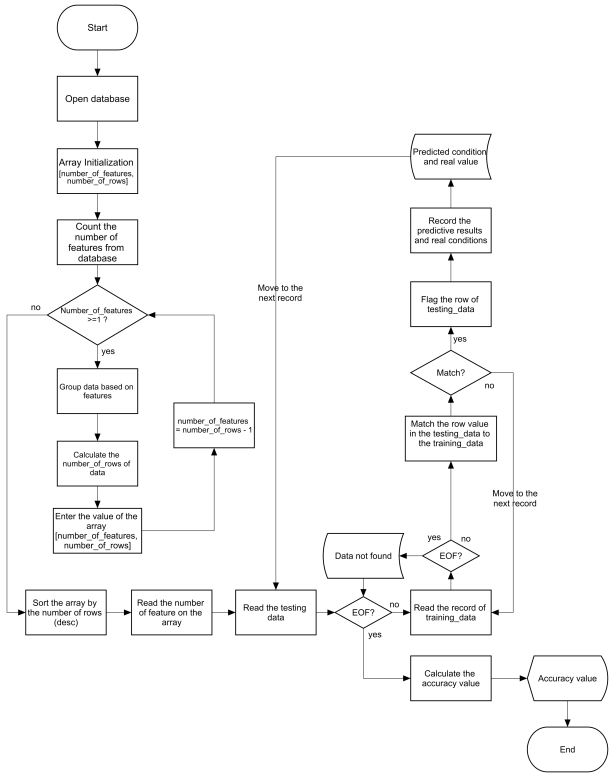

Fig. 3: The classification rules proposed

In Fig. 3, it shows that the classification process begins with feature selection. Then make the selection of features in the data set by getting the order of the most dominant feature in determining the formation of data. The feature is eliminated based on the feature weighting sequence from the lowest. The rules are subsequently transformed into decision trees and used to classify target class determinations based on attributes such as date, month, time, day, weather, event, and road name.

Initially, the prediction model built with the TR-P method was only able to produce an accuracy of 36.12% with six classes of road density. The reason for this low accuracy is the discovery of a target class that is of the null value or has no decision in the target class. The expected road traffic density classes generated after the classification process can be seen in Table 2.

Table 2: Road traffic density classes

No.	Road traffic conditions
1.	*lancar*
2.	*ramai lancar*
3.	*padat*
4.	*padat merayap*
5.	*tersendat*
6.	*macet*

Table 2 shows the six types of road traffic density conditions that have been reported through twitter, including *lancar* (fluent), *ramai lancar* (smooth), *padat* (congested), *padat merayap* (crawl), *tersendat* (stagnant) and *macet* (jammed). These level conditions have been aligned with the regulation of The Minister of Transportation Decree No. 14/2006 on road service conditions.

Based on the results of modeling process then the next work focused on data transformation activities that have a goal to increase the value of accuracy in the prediction model built. We have performed five data transformation experiments, and combine them in the next learning process. The experiments included a grouping of the minute values of reported tweets, a grouping of the days (Monday, Tuesday and so on), a grouping of the months (January, February and so on), a grouping of the dates, and the aggregating of connected roads. The experiments performed and combined with each other resulted in sixteen data sets for subsequent training activities. The experiments up to this stage have succeeded in increasing the accuracy from 36.12% to 43.07%

We also attempted to overcome the predicted failure of the appearance of null values in the road density class. The refinement of the classification method is focused on the analysis of null values on predicted class results declared to be false by the classifier. To overcome the predicted failure of a null value in the traffic density class, we have performed a replacement with the class value that appears at most. The step improvement of the TR-P method we have called with TR-P Advanced.

Another effort to improve the performance of the prediction system is to develop the classifier, among others, applying the Boosting method that works to re-classify the wrongly predicted results. This method uses a technique in AdaBoost that has been known before. The basic concept of the Boosting method is to calculate the weight distribution on the training dataset and modify the distribution on each iteration of the algorithm. We have applied the stacking and voting method to the classifier with the schema as shown in Fig. 4.

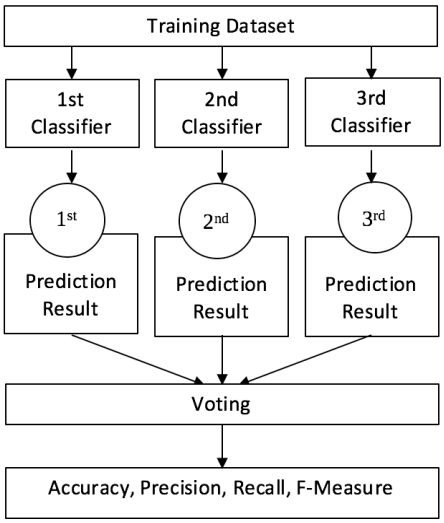

Fig. 4: Classification scheme by Stacking and Voting method

In Fig. 4, it shows that we have combined three classifiers. The first classifier is a combination of the TR-P method with AdaBoost, TR-P method with the C.45 algorithm as the second one and TR-P method with Naïve Bayes as the third. The final prediction decision with this technique is obtained from the voting process of the results of each classifier. This technique does not guarantee better accuracy because the wrong predictive value of the classifiers can be the most value in the voting process. The comparison of training accuracy data of all classifiers using the k-fold method, seen in Table 3.

Table 3: The final accuracy of the data training

No.	Dataset Variant	Training Accuracy				
		TR-P + Adaboost	TR-P + DT	TR-P + NB	Voting	TR-P Advanced
1.	Dataset O	40.63%	40.80%	40.80%	40.80%	44.30%
2.	Dataset A	49.83%	49.83%	49.83%	49.87%	49.97%
3.	Dataset B	41.06%	44.72%	41.67%	42.00%	44.70%
4.	Dataset C	38.71%	43.36%	42.22%	42.20%	43.37%
5.	Dataset D	41.21%	41.04%	41.48%	41.27%	44.93%
6.	Dataset E	40.24%	41.27%	41.41%	40.27%	44.63%
7.	Dataset AB	38.90%	38.87%	38.90%	39.00%	38.97%
8.	Dataset AC	48.32%	48.42%	48.42%	48.07%	48.07%
9.	Dataset AD	38.92%	38.98%	38.88%	39.17%	39.20%
10.	Dataset AE	50.69%	51.79%	51.99%	50.40%	50.73%
11.	Dataset ABC	42.78%	42.88%	42.92%	43.27%	43.33%
12.	Dataset ABD	34.97%	49.70%	45.22%	46.50%	42.80%
13.	Dataset ABE	51.19%	52.06%	53.17%	50.87%	51.03%
14.	Dataset ACD	45.82%	45.85%	45.85%	45.37%	45.37%
15.	Dataset ACE	48.34%	51.15%	51.15%	49.30%	49.30%
16.	Dataset ADE	51.79%	52.16%	52.16%	51.40%	51.47%

Table 3 shows every classifier have quite different accuracy improvement values. The classifier that has adopted the TR-P Advanced method has a significant increase in accuracy when compared to other classifiers.

4. Model Testing

Based on the modeling results that have been done, we completed the testing process with different data during the learning process. The amount of data used in the testing process is as many as 2171 rows of data. The accuracy can be seen in Table 4.

Table 4: The accuracy of the proposed model testing

No.	Dataset Variant	Testing Accuracy				
		TR-P + Adaboost	TR-P + DT	TR-P + NB	Voting	TR-P Advanced
1.	Dataset O	32.80%	34.96%	34.96%	34.96%	42.93%
2.	Dataset A	38.60%	40.35%	40.35%	40.35%	40.99%
3.	Dataset B	22.62%	42.33%	42.33%	42.33%	42.33%
4.	Dataset C	25.79%	43.07%	43.07%	43.07%	43.07%
5.	Dataset D	20.91%	35.01%	42.98%	20.91%	42.98%
6.	Dataset E	19.71%	34.50%	34.50%	34.50%	42.51%
7.	Dataset AB	35.01%	37.59%	37.40%	37.40%	37.59%
8.	Dataset AC	38.88%	40.26%	40.26%	40.26%	40.26%
9.	Dataset AD	35.70%	32.66%	36.53%	35.70%	36.53%
10.	Dataset AE	35.42%	35.42%	35.70%	35.42%	36.48%
11.	Dataset ABC	37.82%	38.88%	38.69%	38.88%	38.88%
12.	Dataset ABD	24.27%	40.90%	28.24%	24.27%	40.90%
13.	Dataset ABE	37.17%	37.17%	38.00%	37.17%	38.23%
14.	Dataet ACD	37.54%	40.53%	40.53%	40.53%	40.53%
15.	Dataset ACE	41.92%	42.01%	42.01%	42.01%	42.01%
16.	Dataset ADE	33.67%	34.59%	34.59%	34.59%	34.82%

On Table 4, it shows that the stability of the accuracy is more visible in the classifier type using the Advanced TR-P method with an average value of 40.07%. Nevertheless, we have made a final effort to increase the value of accuracy by engineering the target class of predicted results. The class of road density, originally text-based then grouped and visualized in color for each class of road density: red, orange, yellow and green.

The modification of the target class into several colors also aims to make the predicted results more easily understand by the users. The change of the density class into four colors (red, orange, yellow, and green), has achieved an accuracy of 91.34%. Meanwhile, the re-classification of road density classes using three color levels (red, yellow, green) obtained 74.48% accuracy.

5. Conclusion

The conclusion that can be obtained from this study indicates that the task of processing text is an important phase in the formation of prediction system of road traffic density based on twitter data. Unique Twitter text data types require precise and accurate text processing to obtain attribute values used in classification activities.

The major challenges that have been solved in this study are the identification of specific roads in an area and how the text data has been successfully classified to construct an accurate traffic density prediction system. The next challenge that has been successfully solved is how to present predicted results to a form that is more easily understood by users. The class of road density originally text-based with an elusive sequence of density levels becomes easier to use when road density classes are visualized in colors.

An interesting fact of the completed road traffic density modeling was found in the final test phase that the accuracy of the performance of all classifiers has been able to exceed the 50% threshold, and this is in line with the expected research objectives. ***

References

[1] A. Fernández-Caballero, F. J. Gómez, and J. López-López, 'Road-traffic monitoring by knowledge-driven static and dynamic image analysis', *Expert Systems with Applications*, vol. 35, no. 3, pp. 701–719, Oct. 2008.

[2] M. Jokela, M. Kutila, J. Laitinen, F. Ahlers, N. Hautière, and A. I. Road, 'Optical Road Monitoring of the Future Smart Roads – Preliminary Results', *International Journal of Computer and Information Engineering*, pp. 502–507, 2007.

[3] R. Bauza, J. Gozalvez, and J. Sanchez-Soriano, 'Road traffic congestion detection through cooperative Vehicle-to-Vehicle communications', *IEEE Local Computer Network Conference*, pp. 606–612, Oct. 2010.

[4] D. Rosenbaum, J. Leitloff, F. Kurz, O. Meynberg, and T. Reize, 'Real-Time Image Processing For Road Traffic Data Extraction From Aerial Images', in *ISPRS TC VII Symposium – 100 Years ISPRS*, 2010, vol. XXXVIII, pp. 469–474.

[5] S. K. Endarnoto, S. Pradipta, A. S. Nugroho, and J. Purnama, 'Traffic Condition Information Extraction and Visualization from Social Media Twitter for Android Mobile Application', *International Conference on Electrical Engineering and Informatics*, no. July, 2011.

[6] R. Kosala and E. Adi, 'Harvesting Real Time Traffic Information from Twitter', *Procedia Engineering*, vol. 50, no. Icasce, pp. 1–11, Jan. 2012.

[7] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, 'Real-Time Detection of Traffic from Twitter Stream Analysis', *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2269–2283, 2015.

[8] S. Bhosale and S. Kokate, 'Traffic Detection Using Tweets on Twitter Social Network', *International Journal of Science and Research (IJSR) ISSN (Online Index Copernicus Value Impact Factor*, vol. 14611, no. 12, pp. 2319–7064, 2013.

[9] S. O. I and N. Salim, 'Opinions From Tweets As Good Indicators Of Leadership And Followership Status', *ARPN Journal of Engineering and Applied Sciences*, vol. 10, no. 3, pp. 1045–1050, 2015.

[10] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom, and P. Chaovalit, 'Social-based traffic information extraction and classification', *2011 11th International Conference on ITS Telecommunications*, pp. 107–112, Aug. 2011.
[11] M. Firdhous, 'Automating Legal Research through Data Mining', *International Journal of Advanced Computer Science and Applications*, vol. 1, no. 6, pp. 9–16, 2010.
[12] B. Sowan and H. Qattous, 'A Data Mining of Supervised learning Approach based on K- means Clustering', *International Journal of Computer Science and Network Security*, vol. 17, no. 1, pp. 18–24, 2017.
[13] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, 'Short Text Classification in Twitter to Improve Information Filtering', *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10*, p. 841, 2010.
[14] H. Liu, B. Luo, and D. Lee, 'Location Type Classification Using Tweet Content', *2012 11th International Conference on Machine Learning and Applications*, pp. 232–237, Dec. 2012.
[15] I. A. Khan, J. Woo, J. Seo, and J. Choi, 'Text Mining : Extraction of Interesting Association Rule with Frequent Itemsets Mining for Korean Language from Unstructured Data', vol. 10, no. 11, pp. 11–20, 2015.

Authors

Arief Wibowo, He received his Bachelor degree in Information System (S.Kom), and Master degree in Computer Science (M.Kom) from Universitas Budi Luhur, Indonesia. Currently, he is a lecturer at Department of Information System, Faculty of Information Technology, Universitas Budi Luhur. His research interests include User Behaviour of Information Technology & Information System, Knowledge Management, and Data Mining. He is a student of Doctoral Program in Computer Science of Universitas Gadjah Mada, Indonesia.

Edi Winarko, He received his Undergraduate degree in Statistics (Drs.) from Universitas Gadjah Mada, Indonesia, Master degree in Computer Sciences (M.Sc) from Queen's University of Canada, and Doctoral degree in Computer Sciences (Ph.D.) from Flinders University, Australia. He is a lecturer at the Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada. His research interests include Data Warehousing, Data Mining, and Information Retrieval.

Azhari, He received his Undergraduate degree in Statistics (Drs.) from Universitas Gadjah Mada, Master degree in Computer Sciences (MT.) from Institut Teknologi Bandung, and Doctoral degree in Computer Sciences (Dr.) from Universitas Gadjah Mada. He is a lecturer at the Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada. His research interests include Software Engineering, Project Management, and Intelligent Agent.

***